## 4.4 Knowledge discovery in a collaborative workflow system

Due to the recent advances in statistical learning methods, various machine learning and data mining techniques have been applied in many science domains, such as biology, physics, informatics, to name a few, and proven their successes in discovering previously hidden knowledge from massively collected data. However, it will be a significant challenge to apply such techniques for knowledge discovery and data mining in the proposed highly collaborative and multi-domain experimental system designed for an expansive international collaboration due to the high level of complexity in collecting and managing both the data generated during the experiments and the metadata associated with the analysis workflows. In addition, those data mining and machine learning processes should be performed under the various types of time and space constraints and assure quality of accuracy.

Our goal with this aspect of research is to solve mainly the following data mining problems: i) discovering and mining expert knowledge from the workflow provenance data and ii) data access pattern recognition aimed at increasing data pre-caching chances to reduce data acquisition latency and to keep up with a quickly evolving set of user requirements. The workflow patterns could feed into our workflow recommendation procedures and the data access patterns can be used to select indexing options for the data and to minimize the data movement in workflow orchestration.

Once the data provenance information is ready, mining the information can be performed in an unsupervised or supervised learning setting depending on how one can evaluate results or quality with prior knowledge or not. Various clustering or classification algorithms (such as model-based clustering, high-dimensional data clustering, pairwise data clustering, support vector machine, etc.) can be used to discover expert knowledge but the solutions should meet the various performance requirements in terms of execution speed and accuracy. Many sequence prediction algorithms for file pre-fetching or webpage pre-caching or association rule mining algorithms (such as Apriori) can potentially be used to access derive knowledge from the users' data access patterns.

The major research effort in this area will be to investigate a data provenance and mining system for collecting the information of cross-domain data flows and data access activities of users in the fusion experimental system. We aim to study the viability of combining this knowledge with externally acquired knowledge about the expertise of the users involved in the collaboration. Using this accumulated and collected provenance information; we will study the viability of predicting data access requirements for the entire set of users.

## 4.5 Extended Modules of ICEE

### 4.5.1 Integrated security

Dynamic workflow control must be executed in a restricted and privileged mode as well as in a collaborative mode. The industry standards such as SSL, PKI, OpenID and SAML would allow the clients and servers to communicate with each other confidentially in a globally distributed collaborative environment where access control is decentralized and managed by different institutions. We plan to leverage the security service design in Earth System Grid Federation (ESGF) where the similar federated authentication and authorization infrastructure needs are present. We will also adopt a similar technology that is commonly found in banking industry for further strict access to the Tokamak workflow control, with One-Time-Password (OTP) technology. OTP would allow further strict validation for the privileged users from the distributed environment to the

Tokamak workflow control, and we will be using mobile platform to deliver the OTP. We will integrate the following components in the framework to serve the system authentication and authorization requirements:

- Authentication Service responsible for validating user's authentication information by allowing OpenID to be used in the collaboration.
- Authorization Service responsible for validating authorization information based on the user authentication information, and attaching the user's access control attributes for a trusted client in the form of a digitally signed SAML statement.
- One-Time Password (OTP) over Short Messaging Service (SMS) responsible for delivering OTP over text messaging based on the authorization information for further strict access to the Tokamak workflow control.

### 4.5.2 Training young scientists without the expense of costly experiments

Running experiments on a Tokamak facility such as KSTAR is expensive, and may be too complicated for inexperienced users. Also, experimental opportunities for inexperienced young scientists are rare, and therefore, lack of experiences in young scientists can cost time and efforts in real experiments in production causing faulty experiments with wrong values that may put the facility in jeopardy of destruction from over-heating. Simulations that would replace experiments are expensive to model and build, and simulation capabilities are sometimes not reaching to the experimental capabilities yet. However, knowledge discovery in a collaborative workflow system based on the workflow provenance data, from the above section, enables simulated runs of the experiments based on the previous experimental cases by experts. Workflows and experimental results produce cases that each condition or collection of conditions can be captured as knowledge from the workflow provenance data. The relevant knowledge can be retrieved and used for reasoning new cases for simulated experiments, and adapted and projected for the expected experimental outcomes. The new knowledge can be revised by an expert or through a real experiment, and retained for the next case for simulated experiments. This process based on Case-Based Reasoning (CBR) can help training inexperienced users without the expense of costly experiments. We plan to build an additional CBR-based training system, on top of the workflow framework to help the next generation of scientists.

### 4.5.3 Validating workflows for changing conditions based on machine learning

In addition to helping inexperienced users gain valuable access to the simulated facility, knowledge discovery in the collaborator workflow system can validate real-time or near-line workflow conditions based on the previous experimental cases. We plan to study adaptive CBR for the dynamic workflow validation model to validate the changing conditions, and provide the workflow system with learning capabilities. This additional validation capability in the proposed task of knowledge discovery in the dynamic workflow control framework can prevent accidental control values as well as project the workflow results before executing the workflow.

### 4.5.4 Mobile access to workflow monitoring

The portability and computing power of modern mobile devices is increasing, and there have been integration activities in a few fields such as medicine and defense as well as research activities in interactive mobile workspace and mobile database for geoscience. The mobile device is finding its way in sciences from space exploration to epidemiology as well as moving into our daily activities beyond workstations and laptops.

However, current technical aspect of mobile device usage, even in the mobile friendly science applications, is somewhat limited to the social networking service (SNS) and mere system status checks. In the context of mobile platform extension of the science workspace, several technical studies and exploration should be needed to prepare for coming computing platform changes and for collaborative science efforts.

As tablet mobile computers are replacing traditional computing platforms (e.g. laptops and workstations) in our daily activities, it is inevitable that tablet computers will soon be an integral part of the scientific community. Supporting tablet computing in scientific research will open up a revolutionary horizon in collaborative environments as well as in scientific computing environments. To explore tablet mobile computing support in collaborative environment as the first step, we propose to study the following areas:

1) Remote monitoring and control of the workflow and data analysis;
    a) Mobile distributed access to the monitoring information for the workflow and data analysis, through tablets with Android OS or iOS,
    b) Tablet access to monitor and update the workflow,
    c) Tablet access to the results of the workflow,
    d) Tablet access to analysis results and science discovery.
2) Exploration of mobile tablet computing in science;
    a) Exploring mobile tablet computing as a data collection mechanism for distributed science collaboration,
    b) Exploring mobile tablet computing as a computing resource for data analysis,
    c) Exploring mobile tablet computing as a collaborative tool – for example, collaboration in grouping and associating data, analysis, experiments or workflows through tagging and annotation over portable devices.

Mobile tablet computing will enable researchers to access quickly and easily to the workflows and the results, like on their workstations, and to control the analysis and workflow interactively.